# THE EARTH SYSTEM GRID

## Enabling Access to Multimodel Climate Simulation Data

BY D. N. WILLIAMS, R. ANANTHAKRISHNAN, D. E. BERNHOLDT, S. BHARATHI, D. BROWN, M. CHEN, A. L. CHERVENAK, L. CINQUINI, R. DRACH, I. T. FOSTER, P. FOX, D. FRASER, J. GARCIA, S. HANKIN, P. JONES, D. E. MIDDLETON, J. SCHWIDDER, R. SCHWEITZER, R. SCHULER, A. SHOSHANI, F. SIEBENLIST, A. SIM, W. G. STRAND, M. SU, AND N. WILHELMI

Current data sharing technologies used in a "unified virtual environment" can support the infrustructural needs of the national and international climate community to securely access, monitor, catalog, transport, and distribute petabytes of data.

Climate scientists face a wide variety of practical problems, but there is an overarching need to efficiently access and manipulate climate model data. Increasingly, for example, researchers must assemble and analyze large datasets that are archived in different formats on disparate platforms, and extract portions of datasets to compute statistical or diagnostic metrics "in place." The need for a common virtual environment in which to access both climate model datasets and analysis tools is therefore keenly felt. The software infrastructure to support such an environment must therefore not only provide ready access to climate data, but also must facilitate the use of visualization software, diagnostic algorithms, and related resources.

To this end, the Earth System Grid Center for Enabling Technologies (ESG-CET) was established in 2006 by the Scientific Discovery through Advanced Computing (SciDAC)-2 program of the U.S. Department of Energy (DOE) through the Office of Advanced Scientific Computing Research (OASCR) and the Office Biological and Environmental

**AFFILIATIONS:** WILLIAMS AND DRACH—Lawrence Livermore National Laboratory, Livermore, California; ANANTHAKRISHNAN, FOSTER, FRASER, AND SIEBENLIST—Argonne National Laboratory, Argonne, Illinois; BERNHOLDT, CHEN, AND SCHWIDDER—Oak Ridge National Laboratory, Oak Ridge, Tennessee; BHARATHI, CHERVENAK, SCHULER, AND SU—Information Services Institute, University of Southern California, Marina del Ray, California; BROWN, CINQUINI, FOX, GARCIA, MIDDLETON, STRAND, AND WILHELMI—National Center for Atmospheric Research,* Boulder, Colorado; HANKIN AND SCHWEITZER—NOAA/PMEL, Seattle, Washington; JONES—Los Alamos National Laboratory, Los Alamos, New Mexico; SHOSHANI AND SIM—Lawrence Berkeley National Laboratory, Berkeley, California

*The National Center for Atmospheric Research is sponsored by the National Science Foundation

**CORRESPONDING AUTHOR:** Dean Williams, Lawrence Livermore National Laboratory, Mail Stop: L-103, P.O. Box 808, Livermore, CA 94551-0808
E-mail: williams13@llnl.gov

*The abstract for this article can be found in this issue, following the table of contents.*
DOI:10.1175/2008BAMS2459.1

In final form 14 July 2008

Research (OBER) within the Office of Science (Fig. 1). ESG-CET is working to advance climate science by developing computational resources for accessing and managing model data that are physically located in distributed multiplatform archives.

**ESG IMPACT ON THE CLIMATE COMMUNITY.** The ESG seeks to address these data challenges of climate science. ESG's infrastructure improves research efficiency by enabling rapid access and analysis of even the largest climate datasets. Since ESG's production launch in 2004, users have downloaded over 2 million files totaling more than 470 terabytes (TB) of data (1 TB = $10^{12}$ bytes).

Current data holdings include the extensive multimodel database of the Coupled Model Intercomparison Project 3 [CMIP3; formerly known as the model data for the Intergovernmental Panel on Climate Change (IPCC) Fourth Assessment Report (AR4)]; ESG also serves data from the Cloud Feedback Model Intercomparison Project (CFMIP) and the output data of the Community Climate System Model (CCSM) and the Parallel Climate Model (PCM). Newer to ESG's scope (since 2007) is the data archive of the Carbon-Land Model Intercomparison Project (C-LAMP), in which the effects of embedding different land biogeochemistry (BGC) schemes in the CCSM coupled ocean–atmosphere climate model are being studied (Hoffman et al. 2007). The C-LAMP experimental output data are archived on an Oak Ridge National Laboratory (ORNL) site that is modeled after the ESG CMIP3 database.

When "publishing" model data into an archive, the providers are given access to ESG catalog metadata (data descriptors) so that they can register the appropriate data characteristics (e.g., dataset title, variable names, spatial and temporal boundaries, etc.). This allows the management of all information pertinent to generating, defining, archiving, and retrieving model simulations. The providers also may restrict access to their data, according to various criteria. ESG's long-term goal is to tie the data information process closely to the climate modeling workflow. In this way, the simulation metadata can be ingested routinely in the ESG archive, thereby expediting the data publishing process and minimizing processing errors.

To illustrate the details of the data publishing process, the protocol followed for the CMIP3 archive of multiple coupled ocean–atmosphere model simulation data is illustrative. This process, coordinated by the Program for Climate Model Diagnosis and Intercomparison (PCMDI) at the Lawrence Livermore National Laboratory drew on PCMDI's considerable prior experience in managing data for international climate model intercomparison experiments.

For the CMIP3 intercomparison, which began in 2004, some 20 participating model development groups performed numerous control and climate change scenario experiments prescribed by the IPCC (Meehl et al. 2007). Adherence to a common data format and metadata standard were essential for the success of such an ambitious model intercomparison. The modeling groups thus were required to transform their simulation outputs into the network common data form (netCDF) format and Climate and Forecast (CF) metadata convention.[1] To facilitate this process, PCMDI provided Climate Model Output Rewriter (CMOR; pronounced "Seymour") software, which ensured efficient production of CF-compliant netCDF files.

Once this format conversion was complete, each modeling group transmitted its simulation data to PCMDI via 1-TB disks. After preliminary quality-control checking, PCMDI software engineers used ESG tools to publish all multimodel simulation data into the CMIP3 archive by the end of 2005. Hundreds of users subsequently have accessed these data in the following variety of ways: via File Transfer Protocol (FTP), the ESG data portal, the Live Access Server (LAS), and the Open-source Project for a Network Data Access Protocol (OPeNDAP; see the appendix and Web site information for further details on these technologies).

These ESG-enabled data archives are freely accessible, once a candidate user completes a simple approval process. While mainly research scientists have accessed these data, the CMIP3 user community also includes educators, students, and employees of the United States and other governments. Their analyses of these data for myriad applications have resulted in authorship of hundreds of climate research publications and impacts reports (online at www-pcmdi.llnl.gov/ipcc/subproject_publications.php).

**HISTORY.** Work on ESG began in the year 2000 with the "Prototyping an Earth System Grid" (ESG I) project, supported by the DOE. In this preliminary phase, ESG developed data grid technologies for managing the movement and replication of large datasets, and applied these to an ESG-enabled data

[1] NetCDF is a set of software libraries and machine-independent data formats that support the creation, access, and sharing of scientific data. The CF metadata conventions are designed to promote the processing and sharing of netCDF data files for climate and forecast applications.

browser based on PCMDI's Climate Data Analysis Tools (CDAT) for analysis and visualization. At a 2001 Supercomputing Conference, ESG demonstrated the potential for remotely accessing and analyzing climate data distributed at several sites across the United States, achieving cross-country transfer rates of more than 500 Mb $s^{-1}$ (Chervenak et al. 2003).

While the ESG I prototype demonstrated proof-of-concept capabilities, the 2002 SciDAC-funded Earth System Grid II project, "Turning Climate Datasets into Community Resources" (Bernholdt et al. 2005), transformed concept into practical reality. ESG II efforts focused on the developing technologies pertinent to extracting metadata from netCDF files and catalog services, security through Web-based user registration and authentication, data transport via the OPeNDAP-g protocol that obviates the need for file downloads, and Web portal accessibility to climate data holdings.

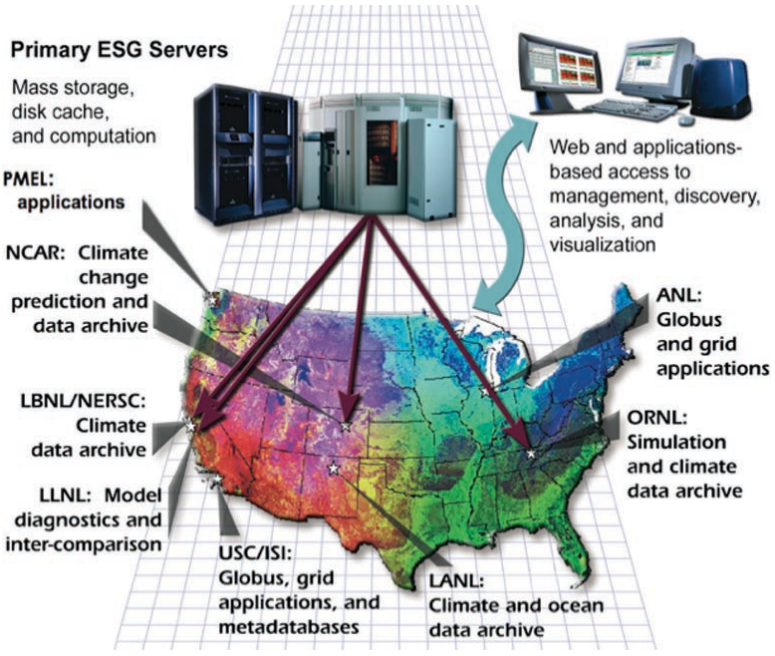


**Fig. 1. The ESG-CET collaboration includes participation from ANL, LANL, LBNL, LLNL, NCAR, ORNL, PMEL, and the USC Information Sciences Institute.**

In order to become more central to the work of climate scientists (Bernholdt et al. 2005), ESG began to distribute CCSM and PCM model data by mid-2004. This first production system brought major advances in model archiving, data management, and sharing of distributed climate data (Allcock et al. 2005; Chervenak et al. 2002, 2006). Originally distributed among just three sites [the National Center for Atmospheric Research (NCAR), Lawrence Berkeley National Laboratory (LBNL), and ORNL; see glossary], this system now supports over 10,000 registered international users and manages some 200 TB of data.

The first-generation ESG architecture (Fig. 2) integrated a wide range of grid and standard information technology tools. The ESG portal (at www.earthsystemgrid.org) was the main access point to the system, providing a central location for authentication, authorization, and accounting services. This portal brokered user data requests among the distributed data nodes, and also provided an interface through which authorized providers published data.

In some cases, individual files or groups of files are too large to be transferred via the ESG portal. For such cases, ESG developed a DataMover tool to implement robust large-scale data interaction with Storage Resource Managers (SRMs; Shoshani et al. 2003), and to replicate thousands of files between specified mass storage systems. A client-side version of this tool, DataMover-Lite (DML), automates multifile data transfers from SRMs into client file systems. The DML's user-friendly interface allows easy monitoring of file transfers to a client machine (Fig. 2).

By late 2004, ESG began distribution of climate model data archived at PCMDI that were relevant for the IPCC's AR4. This database subsequently was designated as the CMIP3 multimodel archive to emphasize its continuity with some 15 yr of international coupled model intercomparisons (e.g., the precursor CMIP1 and CMIP2 efforts). The CMIP3 designation also emphasizes PCMDI's commitment to extend these holdings and to include additional data from current-generation coupled climate models (Meehl et al. 2007).

With its combined data-delivery strategy (via FTP, ESG Web portal, OPeNDAP, or LAS; see the appendix), the CMIP3 archive has distributed more than 420 TB of data to date. The data portal now supports some 2,000 registered analysis projects and manages over 35 TB of data (~80,000 files). By facilitating such widespread data access, ESG also has provided communications channels for CMIP3 data users to suggest enhancements of the accuracy, portability, and performance of current-generation climate models. In response to this concrete user experience, ESG development continues apace.

**FUTURE DIRECTIONS.** In recent years, ESG has entered a new organizational form as the ESG-CET, with funding from DOE's OASCR and OBER. The primary goal of ESG-CET is to generalize the existing system to support archive sites and data types that are more international, broadly distributed, and diverse. A secondary goal is to extend ESG capabilities so that a user can conduct initial analysis of data where it physically resides before downloading the derived analysis products to the client site. ESG-CET views such modalities as essential for promoting widespread use of the petabytes [1 petabyte (PB) = $1 \times 10^{15}$ bytes] of climate data that are potentially available for analysis. ESG-CET thus intends to develop "petascale" data-access capabilities.

In coming years, the ESG-CET will scale up existing capabilities to meet the needs of the following several ambitious scientific projects:

- The North American Regional Climate Change Assessment Program (NARCCAP) will disseminate high-resolution regional climate model data through ESG portals located at both PCMDI and NCAR.
- The Computational Climate End Station (CCES) at the DOE Leadership Computing Facility at ORNL will advance climate science through both an aggressive model development activity and an extensive suite of climate simulations.
- CMIP's Phase 5 (CMIP5) will support the challenging climate data needs of the IPCC's planned Fifth Assessment Report (AR5).

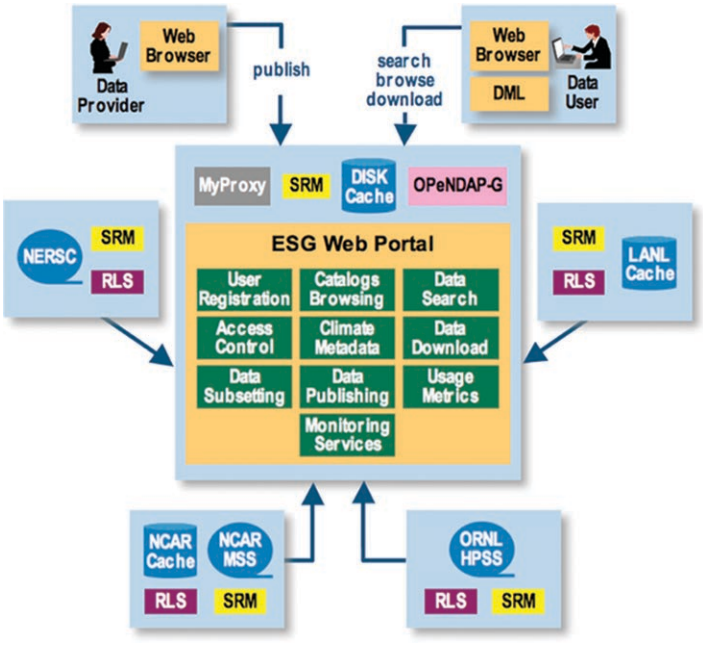


**Fig. 2. Schematic of the first-generation ESG architecture showing the U.S. repositories. Climate model data are located on deep archives at the LBNL National Energy Research Scientific Computing Center (NERSC), the NCAR Mass Storage System (MSS), the ORNL High Performance Storage System (HPSS), or the LANL fast-access rotating disks. Also depicted is a provider publishing data by means of a Web browser and a data user accessing published data via either a Web browser or DML tool. A Replica Location Service (RLS) server at each data node indexes the physical files (or "replicas") available at that site. Online files are served via an Lightweight Authorized HTTP File Server (LAHFS). Files on deep storage are requested and served via an Storage Resource Manager, which retrieves them from the archive and transfers them to a central disk cache, where they are made available by another LAHFS server. All ESG system components are continuously monitored, and system administrators and users are notified whenever a service became unavailable.**

These projects, and especially CMIP5, will drive future development of ESG technologies in order to connect a large number of users with geographically distributed climate model archives, and to provide them with advanced data analysis tools.

Together with its institutional collaborators, ESG-CET will extend its present capabilities in order to supply additional types of climate model data/metadata, to provide more powerful server-side access and analysis services, to enhance interoperability among commonly used climate analysis tools, and to enable end-to-end simulation and analysis workflow (Fig. 3). The following subsections include further details on these plans.

*Institutional collaborators.* Future ESG-CET activities will be framed by relationships with other institutions that share common data-management interests, organized as the Global Organization for Earth System Science Portal (GO-ESSP) consortium. GO-ESSP will develop a common software infrastructure for acquisition and analysis of climate model data. Consortium members that will take leading roles as gateways and/or nodes in the CMIP5 (IPCC AR5) testbed include PCMDI, NCAR, ORNL, and LANL (Fig. 1). Other members that will play a vital role in the CMIP5 effort include the Geophysical Fluid Dynamics Laboratory (GFDL), the British Atmospheric Data Centre (BADC), the World Data Center for Climate (WDCC), and the University

of Tokyo Center for Climate System Research. Because GO-ESSP extends beyond U.S.-based partnerships, it also may need to develop software to accommodate components from the U.K. Natural Environment Research Council (NERC) DataGrid (NDG), the European Union (EU) MetaFor project, and the German C3-Grid initiative.

*Future usage.* Under the current ESG system, a user first accesses and queries a remote database by means of a Web browser, and then retrieves desired data records via the ESG data portal, a DML tool, or a Web "get" ("wget") operation. After downloading these records to the local site, the user usually regrids, reduces, and/or further analyzes the data.

This process often requires many data movements that can overtax network, storage, and computing resources. With the next-generation ESG architecture, the user instead will browse, search, and "discover" (i.e., determine the properties of) distributed data on remote sites. These may include "nontraditional" data products (e.g., biogeochemical and dynamical vegetation variables simulated by CMIP5-coupled climate–carbon cycle models). The user then will be able to regrid and analyze the desired data "*in place*" before downloading them to the local site. This approach will place new data-management demands on ESG hosting sites, but will allow scientific issues, rather than the organization and movement of data, to receive primary attention (Fig. 4).

In future ESG services, the existing Web portal capabilities will be augmented by applications to streamline data download, as well as provide powerful analysis and visualization capabilities. For example, it will be feasible to use popular climate analysis and visualization tools (e.g., CDAT, NCL, GrADS, Ferret, IDL, and MATLAB; see the appendix) directly within the ESG system.

*Functional specification and architecture design.* It is anticipated that computer processing capabilities of the order of $10^{15}$ floating-point operations per second ("petaflops") will be the norm by the year 2010. In order to meet these petascale computational needs, the future ESG architecture must allow for the networking of a large number of distributed sites with varying capabilities. Such "federation" implies that users will have to authenticate only once in order to gain access to data across multiple systems and institutions. To accomplish this objective, the future ESG architecture will be based on three tiers of data services (Fig. 5).

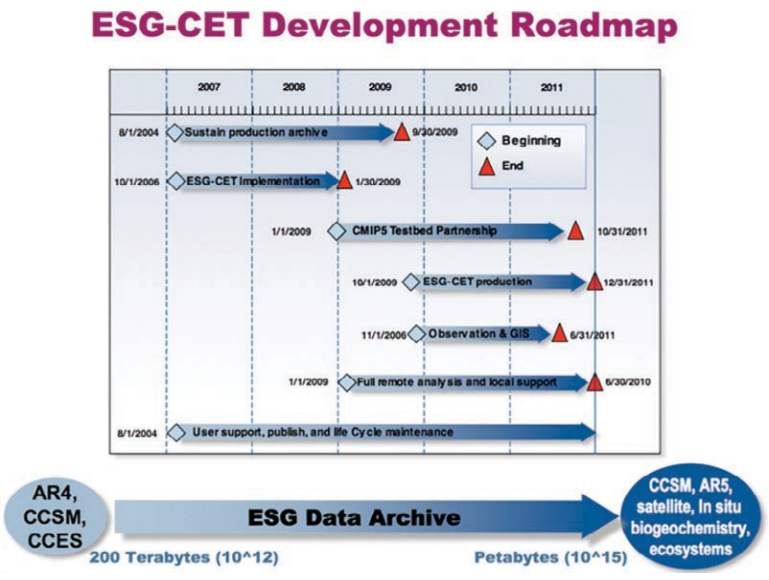


**Fig. 3. A high-level ESG-CET "roadmap" showing the planned evolution of the ESG system from terascale to petascale data management. Also shown are the scientific data management and analysis requirements in relationship to the ESG development timeframe. Note that a distributed testbed for CMIP5 (IPCC AR5) needs to be in place by early 2009.**

Tier 1 services will operate across the entire ESG-CET federation. These include user registration and management, common metadata and notification services to communicate data changes, and global monitoring services to detect data problems. Because all ESG-CET sites will share a common database, a user will be able to find data of interest throughout the entire federation, independent of the site where a data search is initiated.

However, access to specific datasets and related resources will still require approval by the data "owners." Tier 2 data services will comprise multiple ESG gateways that manage limited access to specified data (e.g. the CMIP5 database). Such gateway-deployed services will include the user interface for searching and browsing metadata, for requesting data products (including analysis and visualization tools), and for orchestrating complex workflows. Because the relevant software will require considerable expertise to maintain, tier 2 gateways will be monitored directly by ESG-CET engineers.

Tier 3 will include the actual data holdings and the services used to access these data, which will reside

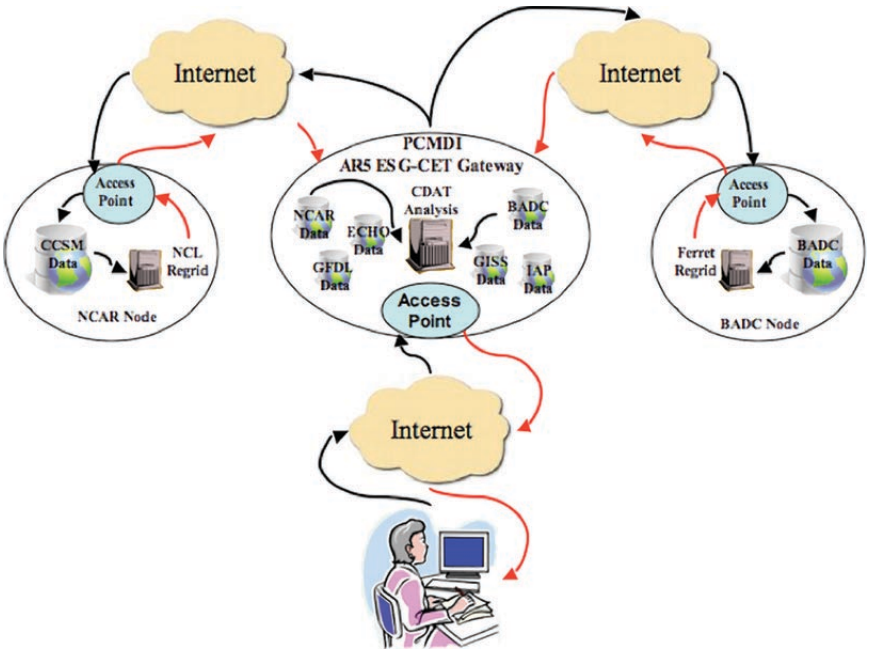


**Fig. 4. In this example a user searches the ESG-CET portal from a remote site. Required data are found at the distributed data node sites [e.g., the NCAR deep storage archive, and the British Atmospheric Data Centre (BADC) fast access disk]. Using popular climate analysis tools (e.g., CDAT, Ferret, NCL), the user regrids the data where they physically reside before transferring the reduced data subset to the PCMDI gateway, where further intercomparison diagnostics are performed on the disparate datasets, and the desired products then are returned to the user's platform.**

on ESG nodes. Tier 3 typically will host the services needed to publish data to ESG, and to execute data-product requests made through an ESG gateway that may serve data requests to many associated nodes: for example, more than 20 institutions are expected to operate ESG nodes for the CMIP5 database. Because personnel with varying levels of expertise will operate ESG nodes, the tier 3 software will come with extensive documentation.

*Component design.* The components designed by ESG-CET will help solve the challenges posed by petascale data archives. An in-depth description of relevant technologies follows.

*Metadata.* Metadata lies at the heart of other major components, especially the search and browse facilities and the publishing system. Thus, metadata design is a priority for ESG-CET.

Current ESG software focuses on metadata for gridded datasets generated from climate model simulations. The next-generation version will expand the scope of data to related subject areas, such as model-based assimilations of observations and predictions of climate change impacts. ESG-CET also will support both derived and virtual datasets. [Derived datasets are products resulting from transformation of one or more "raw" datasets. Virtual datasets have all the properties of traditional data except that they lack location information. For example, a virtual dataset can be generated from a hyperslab request specifying data only in the United States covering a specified time period. Server-side processing to generate derived statistics on the virtual dataset can continue the process further (see Fig. 4).]

In addition, ESG is beginning to design a new search capability based on the concept of "faceted classification" that assigns multiple classifications to an object, allowing these classifications to be ordered in different ways (Adkisson 2003). The user will see search terms and categories that apply within the current context, and thus will be able to avoid queries that would return empty results. Similarly, organizing metadata around facets will provide important flexibility, because the categories can be updated without impacting the rest of the system.

ESG-CET also is working with related metadata projects to ensure consistency with emerging community standards. For example, members of the Earth System Curator (ESC) and MetaFor projects are participating in the design process, and ESG-CET is exploring how the respective metadata schemas intersect. Because both ESC and MetaFor emphasize the viewpoint of the data producer, they have developed schemas that allow a rich description of the structure of models and model components; in contrast, ESG-CET takes the viewpoint of the end user, who is typically more concerned with the scientific aspects of the simulation. The union of these metadata schemas thus will make for a richer and more comprehensive database, and will ensure that ESG can interface with data/metadata derived from ESC and MetaFor.

*Federation.* For ESG-CET, a major challenge in implementing a petascale-distributed architecture is the design of a federated system that allows participating sites to publish datasets and their associated metadata. ESG-CET's envisions a single master metadata catalog that will be hosted at the tier 1 architecture layer (Fig. 5). While this master catalog may be deployed at a particular ESG gateway (e.g., NCAR, ORNL, or PCMDI), all metadata updates will be registered in replica catalogs at each gateway node. Maintaining multiple metadata catalogs helps prevent network bottlenecks and ensures that user queries can be answered even if the master catalog is temporarily unavailable. ESG-CET also is designing a metadata "harvesting" feature whereby the master metadata catalog will be updated whenever data and associated metadata are published at a tier 3 node.

*Security.* Maintaining the security of data and resources is crucial, but this should not place an undue burden on data users and administrators. A practical security protocol is to require only a single sign-on (SSO) in order for a user's browser or client software to gain access to distributed data. SSO will allow the security function of the ESG portal to be split among multiple servers while users authenticate only within their home domain.

*Product services.* ESG-CET serves users having a broad range of experience and needs, from numerical modelers seeking "raw" data to policy analysts who want only an overview of climate model simulations. Certain users thus require products that will be derived via server-side reductions of large data holdings, and ESG-CET is developing usage scenarios to better meet such needs.

For example, the architecture of the Live Access Server (LAS) is being adapted to convert raw data into analysis products, such as statistical summaries and visualization types (e.g., one- to three-dimensional plots with customized contour levels, color palettes, etc.), as appropriate for a range of users. ESG-CET is working with the SciDAC-2 Center for Enabling Distributed Petascale Science (CEDPS; Baranovski et al. 2007) to develop methods for efficiently processing such requests in a parallel computing environment.

*Querying and browsing.* A new interface through which Web portal and client applications will query and browse data from distributed archives is also being developed. ESG-CET will borrow design ideas for data querying from the Open Geospatial Consortium (OGC), an international industrial consortium that defines specifications for such interfaces. Data search results will be returned as Thematic Realtime Environmental Distributed Data Services (THREDDS), a tree-like data structure that represents a hierarchy of datasets in Extensible Markup Language (XML). In the future, THREDDS XML will be generated dynamically (replacing the current static ESG catalog structure), to allow users to browse data in different ways, for example, by model, numerical experiment, climate variable, etc.

*User interface.* Future interactive access of ESG data archives will probably come via a Web browser and/or an application-based interface. ESG-CET is investigating how to integrate the LAS user interface, as well as interfaces offered by other technologies that would allow for the incorporation of more dynamic interaction without the need to change the basic framework of the ESG portal. As the design of the new user interface evolves, however, this approach will need to be reevaluated.

*Publishing.* To publish ESG data, a provider will need to stage a set of model runs, extract the necessary metadata, and check to ensure that the metadata

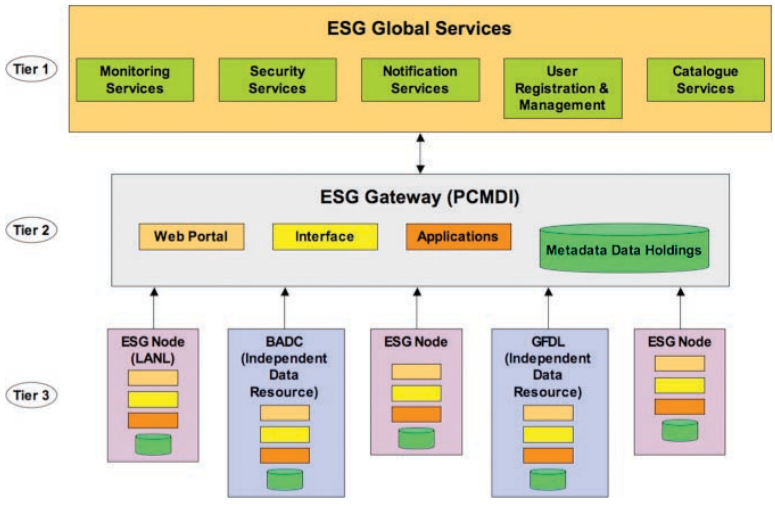


**Fig. 5. Tiered ESG-CET architecture showing the trilevel data services and one of the initial ESG Gateways specific to the CMIP5 (IPCC AR5) application. Initially, three ESG Gateways are planned: one at PCMDI focused on the CMIP5 (IPCC AR5) needs, one at ORNL to support the Computational Climate End Station project, and one at NCAR to serve the CCSM and PCM model communities.**

meet both ESG and project-specific standards. Metadata that are not extractable from the raw model data will have to be added manually. The data provider also will need to specify access privileges for the datasets and obtain publishing privileges by authenticating with a gateway. Every published file or other data aggregation also will receive a unique identifier that remains if it is replicated on a different site. It will be feasible to code this entire publishing process in a script, and to fully document the provenance (origin and history of subsequent owners) of the published data.

**SUMMARY.** The goal of the Earth Systems Grid (ESG) is to catalog and widely publish distributed climate data so as to make it easily accessible to an international community of potential users. The ESG Center for Enabling Technology (ESG-CET) is guiding the evolution of ESG to meet the future petascale climate data demands: deploying server-side analysis services, enhancing interoperability of diverse climate analysis tools and Web portals, and enabling end-to-end workflow.

ESG presently is a large distributed system with primary access offered via two Web portals: one for general climate data and another dedicated to CMIP3 research activities. Current ESG operational services include provisions for metadata and security standards; data transport, aggregation, and subsetting; and monitoring of system and services usage. The ESG-CET team is designing new capabilities to allow climate users to access data products via spinning disk or tertiary storage, to transport data from one site to another; and to increase workflow efficiencies. It is also anticipated that the technologies and capabilities developed in ESG-CET will impact other Scientific Discovery through Advanced Computing (SciDAC-2) applications in the fields of astrophysics, molecular biology, and materials science.

**ACKNOWLEDGMENTS.** The authors extend special thanks to Tom Phillips of PCMDI for his editorial assistance. The work of ESG is supported by the SciDAC program of the U.S. Department of Energy through its Office of Science Advanced Scientific Computing Research and Biological and Environmental Research. Affiliations of participating ESG team members include the following institutions: Argonne National Laboratory (ANL) is managed by University of Chicago Argonne LLC under Contract DE-AC02-06CH11357. Information Sciences Institute (ISI) is a research institute of the Viterbi School of Engineering at the University of Southern California (USC). Lawrence Berkeley National Laboratory (LBNL) is managed by the University of California for the U.S. Department of Energy under Contract DE-AC02-05CH11231. Lawrence Livermore National Laboratory is managed by the Lawrence Livermore National Security, LLC for the U.S. Department of Energy under Contract DE-AC52-07NA27344. Los Alamos National Laboratory (LANL) is managed by Los Alamos National Security, LLC for the U.S. Department of Energy under the Contract DE-AC52-06NA25396. National Center for Atmospheric Research (NCAR) is managed by the University Corporation for Atmospheric Research. Oak Ridge National Laboratory (ORNL) is managed by University of Tennesse—Battelle, LLC for the U.S. Department of Energy under Contract DE-AC-05-00OR22725. Pacific Marine Environment Laboratory (PMEL) is under the National Oceanic and Atmospheric Administration's line office of Ocean and Atmosphere Research, lies within the U.S. Department of Commerce.

## APPENDIX: GLOSSARY.

| | |
|---|---|
| ANL | Argonne National Laboratory, sponsored by the DOE (www.anl.gov/) |
| AR4 | Fourth IPCC Assessment Report, published in 2006 (www.ipcc.ch/ipccreports/ar4-wg1.htm) |
| AR5 | Fifth IPCC Assessment Report, with publication planned in 2011 (www.mnp.nl/ipcc/) |
| BGC | Biogeochemistry, a systems science involving study of the chemical, physical, geological, and biological processes of the natural environment (http://en.wikipedia.org/wiki/Biogeochemistry) |
| CCES | Climate Science Computational End Station, an ORNL data storage archive (www.nccs.gov/leadership-science/climate/climate-science-computational-end-station-development-and-grand-challenge-team/) |
| CCSM | Community Climate System Model, and related data archive maintained by ESG (www.earthsystemgrid.org/) |
| CDAT | Climate Data Analysis Tools, developed by PCMDI (www-pcmdi.gov/software-portal/cdat/) |
| CEDPS | Center for Enabling Distributed Petascale Science, located at ANL (www.cedps.net/) |
| CF | Climate and Forecast metadata convention, for processing and sharing NetCDF data files (http://cf-pcmdi.llnl.gov/) |

| | |
|---|---|
| CFMIP | Cloud Feedback Model Intercomparison Project (http://cfmip.metoffice.com/) |
| CIM | Common Information Model, a product of MetaFor, a metadata standard promoted by the METAFOR organization for describing climate models and their simulation data (www.dmtf.org/standards/cim/) |
| C-LAMP | Carbon-Land Model Intercomparison Project (www.climatemodeling.org/c-lamp/) |
| Client–server | Relationship between two computer programs, where the client program makes a service request that the server program fulfills (http://en.wikipedia.org/wiki/Client-server) |
| CMIP3 | Coupled Model Intercomparison Project 3, sponsored by the World Climate Research Programme's (WCRP's) Working Group on Coupled Modelling (WGCM), and related multimodel database provided for the IPCC AR4 (www-pcmdi.llnl.gov/ipcc/about_ipcc.php) |
| CMIP5 | Coupled Model Intercomparison Project 5, sponsored by WCRP/WGCM, and related multimodel database planned for the IPCC AR5 |
| CMOR | Climate Model Output Rewriter, produces CF-compliant NetCDF data files (www2-pcmdi.llnl.gov/software-portal/cmor/) |
| Data node | Internet location providing data access or processing (http://en.wikipedia.org/wiki/Node-to-node_data_transfer) |
| DML | DataMover-Lite, client software for transferring data files between mass storage and local file systems |
| DOE | Department of Energy, the U.S. Government entity chiefly responsible for implementing energy policy (www.doe.gov/) |
| ESC | Earth System Curator, a software environment for shareing climate-model data and information (www.earthsystemcurator.org/) |
| ESG | Earth System Grid, a collection of hardware and software resources to facilitate access and exchange of distributed climate data (www.earthsystemgrid.org/) |
| ESG-CET | Earth System Grid-Center for Enabling Technologies, leads the development of ESG software and related technologies (www.earthsystemgrid.org/) |
| Exascale | Computer processing capabilities of order 1018 operations per second (http://en.wikipedia.org/wiki/Bit) |
| Ferret | An analysis tool for gridded and nongridded data (http://ferret.wrc.noaa.gov/Ferret/) |
| FTP | File Transfer Protocol, provides intercomputer data transfer across the Internet (http://en.wikipedia.org/wiki/File_Transfer_Protocol) |
| Gateway | Hardware to route data traffic from a computer workstation to an external network that is serving data (http://en.wikipedia.org/wiki/Gateway) |
| GO-ESSP | Global Organization for Earth System Science Portals, develops software for access and analysis of climate data (http://go-essp.gfdl.noaa.gov/) |
| GrADS | Grid Analysis and Display System, developed by the Center of Ocean–Land–Atmosphere Studies (www.iges.org/grads/) |
| GridFTP | Extension of the standard FTP for use with the ESG (http://dev.globus.org/wiki/GridFTP/) |
| HTML | Hypertext Markup Language, the chief means for creating linked documents on the World Wide Web (www.w3.org/MarkUp/) |
| IDL | Interface Description Language, a data visualization and analysis platform (http://rsinc.com/idl/) |
| IPCC | Intergovernmental Panel on Climate Change, a scientific body of the United Nations, periodically issues assessment reports (ARs) on climate change (www.ipcc.ch/) |
| LANL | Los Alamos National Laboratory, sponsored by the DOE (www.lanl.gov/) |
| LAS | Live Access Server, a portal for remotely accessing and manipulating data (http://ferret.pmel.noaa.gov/Ferret/LAS/home/) |
| LBNL | Lawrence Berkeley National Laboratory, sponsored by the DOE (www.lbl.gov/) |
| LLNL | Lawrence Livermore National Laboratory, sponsored by the DOE (www.llnl.gov/) |
| MATLAB | A numerical computing environment and programming language (http://en.wikipedia.org/wiki/Metadata) |
| Metadata | Data properties, such as their origins, spatiotemporal extent, and format (http://en.wikipedia.org/wiki/Metadata) |

| | |
|---|---|
| MetaFor | Organization developing/promoting the metadata standard CIM (www.cgam.nerc.ac.uk/pmwiki/PRISM/index.php/Main/METAFORPage/) |
| NARCCAP | North American Regional Climate Change Assessment Program (www.narccap.ucar.edu/) |
| NCAR | National Center for Atmospheric Research, sponsored by the National Science Foundation (www.ncar.ucar.edu/) |
| NCL | NCAR Command Language (www.ncl.ucar.edu/) |
| NERC | Natural Environment Research Council, a public body of the United Kingdom (www.nerc.ac.uk/) |
| NetCDF | A machine-independent, self-describing, binary data format (www.unidata.ucar.edu/software/netcdf/) |
| NOAA | National Oceanic Atmospheric Administration, an agency of the U.S. Commerce Department (www.noaa.gov/). |
| OASCR | Office of Advanced Scientific Computing Research under the DOE Office of Science (www.er.doe.gov/ascr/) |
| OBER | Office of Biological and Environmental Research under the DOE Office of Science (www.er.doe.gov/ober/) |
| OGC | Open Geospatial Consortium, comprises more than 300 organizations promoting the development of standards for geospatial content and services (www.opengeospatial.org/) |
| OPeNDAP | Open-source Project for a Network Data Access Protocol, an architecture for data transport including standards for encapsulating structured data and describing data attrubtes (www.opendap.org/) |
| OPeNDAP-g | Equivalent in function to OPeNDAP, but implemented in a distributed computing environment (www.opendap.org/) |
| ORNL | Oak Ridge National Laboratory, sponsored by the DOE (www.ornl.gov/) |
| PB | Petabyte, 1015 byte of information (http://en.wikipedia.org/wiki/Petabyte) |
| PCM | Parallel Climate Model, developed by NCAR scientists, includes coupled atmosphere, ocean, land, and sea ice components (www.cgd.ucar.edu/pcm/) |
| PCMDI | Program for Climate Model Diagnosis and Intercomparison, located at LLNL (www-pcmdi.llnl.gov/) |
| Petascale | Computer processing capabilities of order 1015 operations per second—the expected norm by the year 2010 (http://en.wikipedia.org/wiki/Bit) |
| PMEL | Pacific Marine Environmental Laboratory, sponsored by NOAA (www.pmel.noaa.gov/) |
| POP | Parallel Ocean Program, a three-dimensional ocean circulation model developed by LANL scientists (http://climate.lanl.gov/Models/POP/) |
| SciDAC | Scientific Discovery through Advanced Computing, a program within the DOE Office of Science promoting use of high-performance computers in scientific applications (www.scidac.gov/) |
| SDM | Scientific Data Management is one of six research groups of the High Performance Computing Research Department at LBNL (http://sdm.lbl.gov/) |
| Sever side | Operation(s) performed only by the server in a client–server relationship in computer networking (http://en.wikipedia.org/wiki/Server-side) |
| SSO | Single sign-on, a procedure requiring a user to authenticate only once in order to gain access to the resources of multiple software systems (http://en.wikipedia.org/wiki/Single_sign-on) |
| SRM | Storage resource managers, provide dynamic space allocation and file management of shared ESG storage components (http://sdm.lbl.gov/srm-wg/doc/SRM.v2.2.html) |
| TB | Terabyte 1012 (a trillion) storage bytes (http://en.wikipedia.org/wiki/Terabyte) |
| Terascale | 1012 (a trillion) computer operations per second, the current performance standard (http://en.wikipedia.org/wiki/Bit) |
| THREDDS | Thematic Realtime Environmental Distributed Data Services, or the hierarchical tree-like data structure developed by this organization (www.unidata.ucar.edu/projects/THREDDS/) |
| USC/ISI | University of Southern California's Information Sciences Institute (www.isi.edu/index.php) |
| Web portal | A point of access to information on the World Wide Web (http://en.wikipedia.org/wiki/Web_portal) |

| | |
|---|---|
| Workflow | A sequence of operations, performed by person(s), organization(s), or mechanism(s) (http://en.wikipedia.org/wiki/Workflow) |
| XML | Extensible Markup Language, a general-purpose specification for creating custom hypertext (www.unidata.ucar.edu/projects/THREDDS/) |

## REFERENCES


Adkisson, H. P., 2003: Use of faceted classification. Web design practices. [Available online at www.webdesignpractices.com/navigation/facets.htm.]

Allcock, B., J. Bresnahan, R. Kettimuthu, M. Link, C. Dumitrescu, I. Raicu, and I. Foster, 2005: The Globus Striped GridFTP framework and server. *Proc. of the 2005 ACM/IEEE Conf. on Supercomputing,* Seattle, Washington, ACM/IEEE Computer Society, 54–65. [Available online at http://portal.acm.org/citation.cfm?id=1105819.]

Baranovski, A., and Coauthors, 2007: Enabling distributed petascale science. *J. Phys. Conf. Ser.,* **78,** 012020, doi:10.1088/1742-6596/78/1/012020.

Bernholdt, D., and Coauthors, 2005: The Earth System Grid: Supporting the next generation of climate modeling research. *Proc. IEEE,* **93,** 485–495.

Chervenak, A., and Coauthors, 2002: Giggle: A framework for constructing scalable replica location services. *Proceedings of the 2002 ACM/IEEE Conference on Supercomputing,* IEEE Computer Society Press, 1–17. [Available online at www.globus.org/research/papers.html#giggle.]

——, and Coauthors, 2003: High-performance remote access to climate simulation data: A challenge problem for data grid technologies. *Parallel Comput.,* **29,** 1335–1356.

——, and Coauthors, 2006: Monitoring the Earth System Grid with MDS4. *Proc. of the Second IEEE Int. Conf. on e-Science and Grid Computing (e-Science 2006),* Washington, DC, IEEE Computer Society, 69.

Hoffman, F. M., and Coauthors, 2007: Results from the Carbon-Land Model Intercomparison Project (C-LAMP) and availability of the data on the Earth System Grid (ESG). *J. Phys. Conf. Ser.,* **78,** 012026, doi:10/1088/1742-6596/78/1/012026.

Meehl, G. A., C. Covey, T. Delworth, M. Latif, B. McAvaney, J. F. B. Mitchell, R. J. Stouffer, and K. E. Taylor, 2007: The WCRP CMIP3 multi-model dataset: A new era in climate change research. *Bull. Amer. Meteor. Soc.,* **88,** 1383–1394.

Shoshani, A., A. Sim, and J. Gu, 2003: Grid Resource Management: State of the Art and Future Trends. *Storage Resource Managers: Essential Components for the Grid,* J. Nabrzyski, J. M. Schopf, J. Weglarz, Eds., Kluwer Academic Publishers, 229–348. [Available online at http:// sdm.lbl.gov/~arie/papers/SRM.book.chapter.pdf.]